Resident Characterization of Better-than- and Worse-than-Average Clinical Teaching

Bishr Haydar, M.D., Jonathan Charnin, M.D., Terri Voepel-Lewis, M.S., R.N.,
Keith Baker, M.D., Ph.D.

ABSTRACT

Background: Clinical teachers and trainees share a common view of what constitutes excellent clinical teaching, but associations between these behaviors and high teaching scores have not been established. This study used residents' written feedback to their clinical teachers, to identify themes associated with above- or below-average teaching scores.

Methods: All resident evaluations of their clinical supervisors in a single department were collected from January 1, 2007 until December 31, 2008. A mean teaching score assigned by each resident was calculated. Evaluations that were 20% higher or 15% lower than the resident's mean score were used. A subset of these evaluations was reviewed, generating a list of 28 themes for further study. Two researchers then, independently coded the presence or absence of these themes in each evaluation. Interrater reliability of the themes and logistic regression were used to evaluate the predictive associations of the themes with above- or below-average evaluations.

Results: Five hundred twenty-seven above-average and 285 below-average evaluations were evaluated for the presence or absence of 15 positive themes and 13 negative themes, which were divided into four categories: teaching, supervision, interpersonal, and feedback. Thirteen of 15 positive themes correlated with above-average evaluations and nine had high interrater reliability (Intraclass Correlation Coefficient >0.6). Twelve of 13 negative themes correlated with below-average evaluations, and all had high interrater reliability. On the basis of these findings, the authors developed 13 recommendations for clinical educators.

Conclusions: The authors developed 13 recommendations for clinical teachers using the themes identified from the above- and below-average clinical teaching evaluations submitted by anesthesia residents. (Anesthesiology 2014; 120:120-8)

THE Accreditation Council for Graduate Medical Education requires residency programs to provide trainees the opportunity to evaluate their supervising faculty members. These evaluations should include the "faculty's clinical teaching abilities, commitment to the educational program, clinical knowledge, professionalism, and scholarly activities."[1] Such evaluations are often used to support promotion, decide resident assignments, distribute bonus money, and remediate less-skilled teachers.[2,3]

Evaluations that include constructive feedback may be viewed positively by faculty;[4] however, low-performing faculty may be harmed by constructive feedback.[5] Previous studies have shown that constructive criticism generally improves clinical teaching scores over time,[6,7] whereas teaching scores in the absence of constructive feedback have more mixed results.[6] Ideally, comments would reflect specific and modifiable factors, thus, providing guidance for the faculty to improve their performance. For example, when surgical faculty were scored on their suitability as role models, many of the lowest-ranked faculty made significant gains over a

What We Already Know about This Topic

- Constructive feedback from trainees improves faculty teaching scores
- Although trainees identify behaviors of an ideal teacher, whether they utilize these concepts in providing feedback to faculty is not known

What This Article Tells Us That Is New

- In a two-step process, comments from faculty evaluations over a 2-yr period at one institution were studied to identify themes associated with above- and below-average ratings by trainees
- Thirteen themes were identified, using trainee evaluations, and these fell into four domains associated with outstanding teaching

This article is featured in "This Month in Anesthesiology," page 1A. Corresponding article on page 22. Presented at the International Assembly for Pediatric Anesthesia, October 12, 2012, Washington, D.C., and at the American Society of Anesthesiologists Annual Meeting, October 13, 2012, Washington, D.C.

Submitted for publication November 21, 2012. Accepted for publication February 21, 2013. From the Department of Anesthesiology, University of Michigan, Ann Arbor, Michigan (B.H., T.V.-L.), and Harvard Medical School, Department of Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts (J.C., K.B.).

6-month period.[7] Similarly, anesthesiologists showed gains in their overall teaching scores within a year, when provided with both evaluative scores and feedback.[6]

Using surveys and focus groups,[8,9] residents identified four unique roles of the clinician-teacher: physician, supervisor, teacher, and person. These roles are consistent with the Ulian conceptual model, which encompasses specific teaching behaviors and approaches, attitudes toward the trainee, and interpersonal skills.[10] Trainees expressed preferences for teaching faculty who had clear expertise and up-to-date clinically relevant knowledge, provided appropriate autonomy and supervision, provided formative feedback, provided efficient relevant teaching through discussion, exhibited kindness and sensitivity toward the trainees, and adopted a collegial manner.[8,9] When using general descriptive terms, faculty and residents largely share a single view of what constitutes an ideal clinical teacher.[11] To date, the link between these ideals and residents' thoughts and comments when evaluating their best clinical teachers has not been established. To enhance learning of new material, studies have shown that comparing and contrasting new ideas, topics, or themes can be a more effective and efficient method than simply learning one idea, topic, or theme, and then moving on (massed leaning).[12] Comparing and contrasting themes found in evaluations with high and low teaching scores is, thus, an effective way to identify important differences associated with high and low evaluations. The objective of this study was to identify themes found in resident feedback that characterize better-than- or worse-than-average teaching capacities of anesthesia faculty members. We hypothesize that specific behaviors and characteristics will be associated with better- and worse-than-average teaching scores.

Materials and Methods

Evaluation System

Anesthesia residents at the Massachusetts General Hospital evaluate their clinical teachers on a monthly basis using numerical scores and free text, as has been previously described.[6] The evaluation form has seven assessment categories: overall time spent, clinical supervision, quality of teaching, quantity of teaching, role modeling, and encouragement given to think about the science of anesthesia. Each item is rated on a Likert scale (0–10), with 10 representing the highest score. Free-text comments can be entered in three separate boxes: strengths, areas that need improvement, and additional comments. Teaching scores are calculated by summing up the seven individual scores, and thus, they range from 0 to 70.

Evaluations are submitted in a confidential process, where each resident evaluator's name is replaced by a unique number. Faculty members are given their aggregate teaching scores along with normative data and all their free-text comments. Data are released at 6-month intervals and with a delay to ensure resident anonymity.

Study Design

Our study protocol was approved by the Institutional Review Board of Massachusetts General Hospital (# 2011P-000676) and exempted from review by the University of Michigan (# HUM0005519). We used all the resident evaluations of the faculty, submitted from January 1, 2007 to December 31, 2008. Using unique identifiers, we calculated a mean teaching score that each resident gave to the teaching faculty. We retained only those evaluations that had teaching scores 20% higher or 15% lower than the residents' mean teaching scores in order to obtain a substantial number of evaluations, with comments related to above- and below-average teaching scores.

After reading all comments submitted during the first 6-month period, Dr. Baker developed a list of recurring themes (tables 1 and 2). Next, all evaluations in the research database were independently reviewed by two investigators (Drs. Haydar and Charnin) blinded to the teaching score. They determined the presence or absence of these themes. If a particular comment had no particular theme then the investigator could elect "positive, not otherwise specified" (positive NOS), "negative, not otherwise specified" (negative NOS), or none. They also predicted whether the evaluation was associated with an above-average or below-average teaching score based on the comments alone.

Statistics

We evaluated interrater reliability for each theme, using an Intraclass Correlation Coefficient (ICC) and Cohen Kappa and compared the relationship of each theme with the dichotomized above- and below-average scores, using the two-sided Fisher exact test. We accepted ICCs between 0.40 and 0.59 as fair, between 0.60–0.74 as good, and above 0.75 as excellent agreement.[13] Themes with poor interrater reliability (*i.e.*, ICC <0.40) were excluded from further analyses. We performed linear regressions to assess for collinearity between variables. We then excluded from our predictive model themes which loaded solely on above- or below-average evaluations, as they may cause major errors in logistic regressions.[14] Finally, we used a logistic regression model, where the dichotomous outcome, above-average or below-average evaluation, was regressed on the independent themes. All comparisons were two-sided, and *P* value less than 0.05 was accepted as significant. All analyses were conducted using SPSS (version 20; IBM Corporation, Armonk, NY) or Origin (version 7.5 SR4; OriginLab Corp., Northampton, MA).

Results

From January 1, 2007 to December 31, 2008, 117 residents submitted 9,786 evaluations on 162 faculty

Table 1. Positive Themes in High-scoring Evaluations

Theme		
Teaching		Representative examples
p1	Use of primary literature to support teaching	"His educational points are all supported by journal articles."
p2	Explaining why specific management strategies were used	"Takes time to explain her thought process."
p3	Having education-oriented discussions	"Dr. (redacted) and I had an in-depth discussion of tissue hypoxia…"
p4	Spending adequate time teaching	"Makes a point of teaching every day."
p5	Teaching to the appropriate level of the resident	"…teach(es) to the resident's interests and needs."
p6	Demonstrating an active effort in teaching the resident	"… works hard to teach."
p7	Demonstrating and imparting significant clinical knowledge	"Colleagues in anesthesia and surgery seek his opinion and skills for the most complex cases."
p8	Teaching clinically relevant material	"…teaches practical information."
Supervision		
p9	Allowing a healthy balance between supervision and autonomy	"Very patient and promotes appropriate independence in OR."
p10	Having high expectations of the resident	"had very high expectations of his residents."
p11	Providing support while teaching a new procedure	"Patient with procedures."
p12	Challenging the resident to a better performance	"…challenges residents to examine every element of anesthetic care."
p13	Encouraging the use of new methods or procedures	"Encourages residents to try various methods for anesthesia."
Feedback		
p14	Providing developmental feedback	"Gives helpful hints for improving skills."
Interpersonal		
p15	Treating the resident in a collegial and/or respectful manner	"Willing to discuss, listen, give feedback."

Positive recurring themes identified from review of 6 months of evaluations with high and low teaching scores.
OR = operating room.

members. The mean teaching score overall, was 58.3 with an SD of 10.3. There were 527 evaluations that had both a teaching score 20% above the resident's mean and comments. There were 285 evaluations that had both a teaching score 15% below the resident's mean and comments. We excluded nine of these evaluations for unintelligible comments and 198 (25%) that contained only "positive NOS," "negative NOS," or both, leaving 605 evaluations for analysis. Of these, 61% were above average, whereas 39% were below average. Sixty-seven percent of all evaluations had positive comments (97% of above-average, 21% of below-average), whereas 43% had negative comments (8% of above-average, 97% of below-average). The distribution of above- and below-average teaching scores received by each individual faculty member is presented in figure 1. The distribution of above- and below-average teaching scores assigned by each resident is presented in figure 2. A majority of residents (71%) submitted evaluations that ended up in the research database, and 82% of the teaching faculty was represented. Among these faculty members, 29% received exclusively above-average evaluations and 14% received exclusively below-average evaluations, whereas the majority (56%) received both. Among the residents represented, 59% submitted both above- and below-average evaluations, whereas 24% submitted only above-average evaluations, and 17% submitted only below-average evaluations.

Using only comments, we correctly identified 90.8% of below-average evaluations and 94.6% of above-average evaluations. The interrater reliability between reviewers' classifications into above- and below-average evaluations was excellent with an ICC of 0.95. A high degree of interrater reliability[13] (ICC >0.6) was found for 9 of 15 positive themes and for all negative themes (tables 3 and 4). One positive theme (teaching to the appropriate level of the resident. Table 1, p5), and both "Positive NOS" and "Negative NOS" had very poor reliability, and were excluded from further analysis. Nearly all positive themes had statistically significant associations with above-average teaching scores. Only one positive theme (having high expectations for the resident. Table 3, p10) was not significantly associated with above-average teaching scores. Nearly all negative themes had significant associations with below-average teaching scores. Only one negative theme (providing teaching that is

Table 2. Negative Themes in Low-scoring Evaluations

	Theme	
Teaching		Representative examples
n1	Failing to explain why specific management strategies were chosen	"At times, his rationale for doing things is, 'Because that is how I do it.'"
n2	Spending an inadequate amount of time teaching	"Minimal teaching."
n3	Providing teaching that is overly limited in scope or clinically irrelevant	"Teaches only about esoteric topics that are nonclinical."
Supervision		
n4	Being too rigid or prescriptive in the management of a patient	"Quite rigid. Generally unwilling to try something different from his established practices."
n5	Being too passive or unhelpful during busy or challenging times	"Wasn't helpful in assisting during difficult cases"
n6	Intervening prematurely without involving the resident in a decision or a procedure	"Should learn how to "coach" residents rather than step in and take over the procedure."
n7	Providing insufficient supervision or too little autonomy in the management of the patient	"Tends to micromanage."
n8	Having a low clinical ability as perceived by the resident	"His preops are often incomplete and often fails to discuss the full plan with residents."
Feedback		
n9	Failing to provide developmental feedback	"Didn't give me feedback when asked."
Interpersonal		
n10	Becoming impatient, frustrated, or angry with the resident	"Abrupt with OR staff and families alike. Short-tempered."
n11	Being overly critical of the resident	"Please do not judge residents so harshly, give them a chance."
n12	Adopting an intimidating demeanor, or treating the resident in an overly rude, condescending, or abrasive manner	"… tends to be too intimidating and at times difficult to work with."
n13	Speaking ill of other residents who are not present	"Do not speak negatively about residents to other attendings…"

Negative recurring themes identified from review of 6 months of evaluations with high and low teaching scores.
OR = operating room.

overly limited in scope or clinically irrelevant. Table 4, n3) was not significantly associated with below-average teaching scores. No items were found to be collinear, indicating that each theme was independent of the other.

Our logistic regression analysis demonstrated six positive themes, which were independently associated with above-average evaluations (having education-oriented discussions, spending adequate time teaching, demonstrating an active effort in teaching the resident, allowing a healthy balance of supervision to autonomy, providing support during teaching a new procedure, treating the resident in a collegial and/or respectful manner. Table 3: p3, p4, p6, p9, p11, and p15, respectively). In addition, several positive themes were found only among above-average evaluations, but these themes could not be regressed using binary logistic regression due to their highly skewed distribution. Three of these themes were highly correlated with above-average evaluations using Fisher exact test (explaining why specific management strategies were used, challenging the resident to a better performance, providing developmental feedback. Table 3: p2, p12, and p14, respectively).

Our logistic regression analysis also demonstrated eight negative themes, which were independently associated with below-average evaluations (failing to explain why specific management strategies were chosen, spending an inadequate amount of time teaching, being too rigid or prescriptive in the management of a patient, being too passive or unhelpful during busy or challenging times, intervening prematurely without involving the resident in a decision or a procedure, providing insufficient supervision or too little autonomy in the management of the patient, becoming impatient, frustrated, or angry with the resident, adopting an intimidating demeanor, or treating the resident in an overly rude, condescending, or abrasive manner. Table 4: n1, n2, n4, n5, n6, n7, n10, and n12, respectively). In addition, several negative themes were found only among below-average evaluations, but these themes could not be regressed using binary logistic regression because the themes did not appear in the above-average evaluations (no counts). Three of these themes were highly correlated with below-average evaluations using Fisher exact test (having a low clinical ability as perceived by the resident, being overly critical of the resident, speaking ill of other residents who are not present. Table 4: n8, n11, and n13, respectively).

Due to lower interrater reliability (ICC <0.6) of some of the themes in the logistic regression, we performed a second

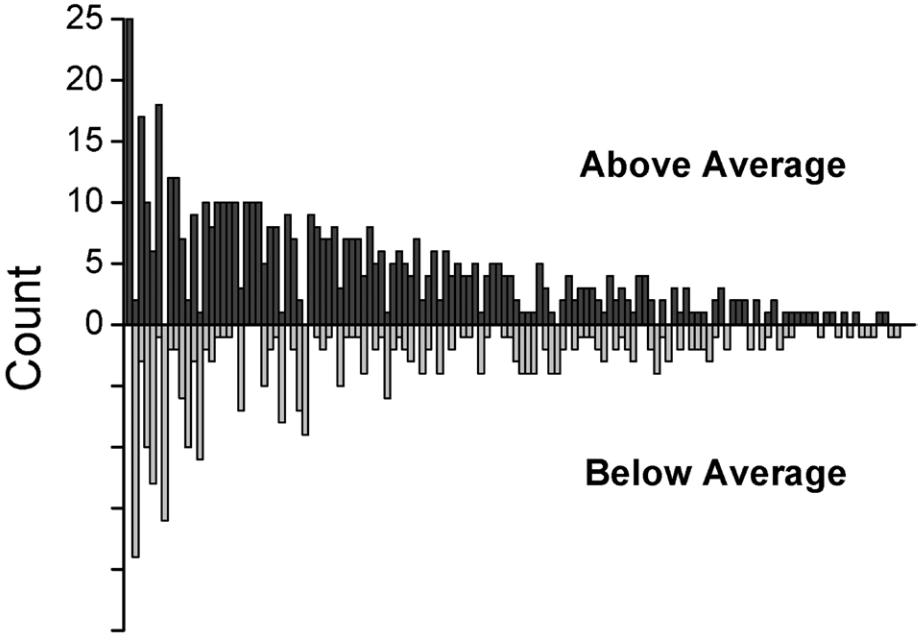

Fig. 1. Counts of above- and below-average evaluations received by each faculty member. Upward *dark gray bars* indicate total number of above-average evaluations received by an individual faculty member. Downward *light gray bars* indicate total number of below-average evaluations received by that same individual faculty member. Faculty members are ordered by the sum total count of above-average and below-average evaluations received. Most faculty members received both above-average and below-average evaluations though a minority received exclusively one type or the other.

and separate logistic regression using only the second rater's themes. To accomplish this regression, we also excluded p13 and n9, as these themes had zero counts for below-average and above-average evaluations, respectively. The regression results, using only the second rater's data, found the same statistically significant associations as the first regression.

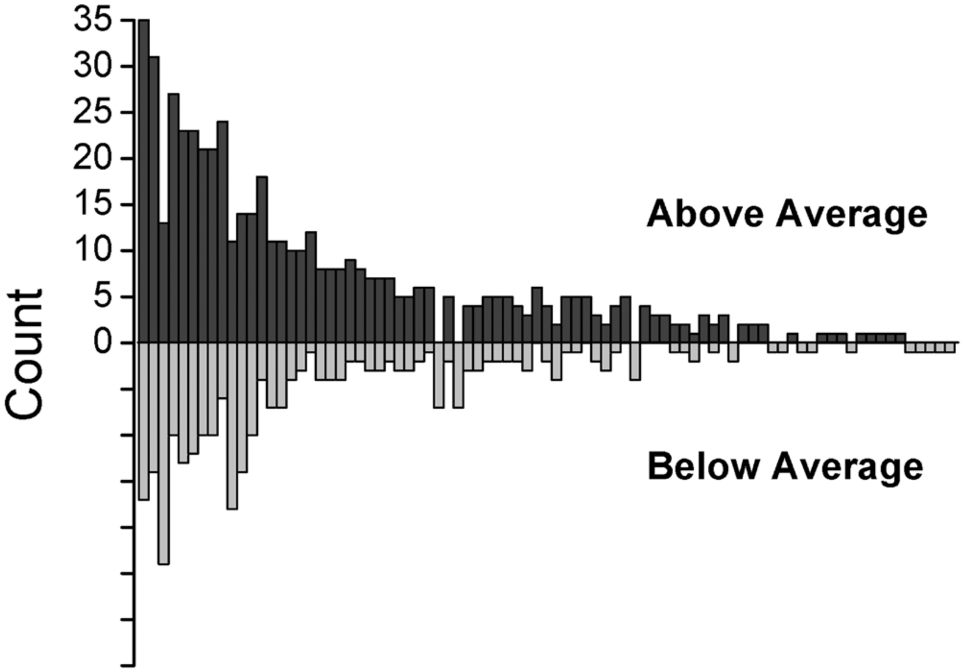

Fig. 2. Counts of above- and below-average evaluations submitted by each resident. Upward *dark gray bars* indicate total number of above-average evaluations submitted by an individual resident. Downward *light gray bars* indicate total number of below-average evaluations submitted that same individual resident. Residents are ordered by the sum total count of above-average and below-average evaluations submitted. Most residents submitted both above-average and below-average evaluations though a minority submitted exclusively one type or the other.

Table 3. Positive Themes: Counts, Reliability, and Regression

Theme	Total	Teaching Score		ICC	Kappa	Fisher Exact	Logistic Regression *P* Value
		Low	High				
Uses primary literature to support teaching (p1)	41	3	38	0.88	0.79	<0.001	0.160
Explaining why specific management strategies were used (p2)	11	0	11	0.66	0.49	0.004	—
Having education-oriented discussions (p3)	21	1	20	0.46	0.30	<0.001	0.012
Spending adequate time teaching (p4)	112	2	110	0.57	0.39	<0.001	<0.001
Teaching to the appropriate level of the resident (p5)	7	0	7	−0.03	−0.01	—	—
Demonstrating an active effort in teaching the resident (p6)	43	2	41	0.59	0.42	<0.001	<0.001
Demonstrating and imparting significant clinical knowledge (p7)	47	12	35	0.75	0.60	0.045	0.179
Teaching clinically relevant material (p8)	18	1	17	0.53	0.35	0.002	0.012
Allowing a healthy balance of supervision and autonomy (p9)	66	7	59	0.91	0.83	<0.001	0.011
Having high expectations of the resident (p10)	5	0	5	0.46	0.30	0.162	—
Providing support while teaching a new procedure (p11)	42	1	41	0.73	0.58	<0.001	0.001
Challenging the resident to better performance (p12)	13	0	13	0.68	0.51	0.002	—
Encouraging the use of new methods or procedures (p13)	18	1	17	0.60	0.43	0.002	0.068
Providing developmental feedback (p14)	30	0	30	0.81	0.68	<0.001	—
Treating the resident in a collegial and/or respectful manner (p15)	92	10	82	0.67	0.50	<0.001	0.014
Positive, not otherwise specified	24	14	10	0.11	0.03	0.057	—

Positive themes, total number of evaluations containing each theme, distribution of themes among above- and below-average evaluations, interrater reliability using Cohen Kappa and ICC (average measures), distribution using two-sided Fisher exact test, and logistic regression significance (when applicable). ICC = Intraclass Correlation Coefficient.

Associations that were not significant in the first regression were not significant when coded by the second rater.

Discussion

Resident evaluations of the teaching faculty can be considered "high-stakes" evaluations with potential implications for promotion, raises, and other professional compensation.[3] Trainee expressions of ideal behaviors and characteristics of teaching faculty fall neatly into the four domains described by the Ulian model: physician, supervisor, teacher, and person.[8,10] Among the themes we evaluated, most address the latter three domains, and many relate to potentially modifiable behaviors. The comments in our study convey diagnostic teaching information because we were able to correctly categorize more than 90% of the evaluations as being above- or below-average based solely on the comments. Residents disproportionately included more positive comments on below-average evaluations than negative comments on above-average evaluations. They also submitted many more above-average evaluations than below-average evaluations. The mean score for faculty teaching was well above the center of the scale, and this indicates grade inflation, as has been seen elsewhere.[6,15] We also encountered below-average evaluations that contained no negative comments (8.6%), and some above-average evaluations that contained no positive comments (3.0%). Characterizing above-average teaching using resident comments may be helpful in identifying why a particular faculty member receives high scores. Using raw scores alone can be problematic because of grade inflation and other independent factors, which influence teaching scores.[16]

Our chosen positive themes provide support for many practices that clinician-educators regard as essential. These include providing both appropriate autonomy and supervision, imparting knowledge, providing developmental feedback, and doing this in a matter that is fitting of a future colleague. Many of our positive themes reflect the trainee's desire to learn and develop as a clinician. Importantly, when residents expressed satisfaction with faculty feedback, the teaching scores were always above-average. Providing constructive feedback is considered to be an essential element of clinical teaching by the Accreditation

Table 4. Negative Themes: Counts, Reliability, and Regression

		Teaching Score					Logistic
Theme	Total	Low	High	ICC	Kappa	Fisher Exact	Regression *P* Value
Failing to explain why specific management strategies were chosen (n1)	17	15	2	0.86	0.75	<0.001	0.002
Spending an inadequate amount of time teaching (n2)	74	70	4	0.90	0.82	<0.001	<0.001
Providing teaching that is overly limited in scope or clinically irrelevant (n3)	5	4	1	0.72	0.57	0.081	0.117
Being too rigid or prescriptive in the management of a patient (n4)	13	12	1	0.73	0.57	<0.001	0.003
Being too passive or unhelpful during busy or challenging times (n5)	20	19	1	0.79	0.66	<0.001	0.004
Intervening prematurely without involving the resident in a decision or a procedure (n6)	14	12	2	0.81	0.68	<0.001	0.017
Providing insufficient supervision or too little autonomy in the management of the patient (n7)	83	76	7	0.75	0.60	<0.001	<0.001
Having a low clinical ability as perceived by the resident (n8)	7	7	0	0.71	0.55	0.001	—
Failing to provide developmental feedback (n9)	6	5	1	0.80	0.66	0.037	0.084
Becoming impatient, frustrated, or angry with the resident (n10)	16	14	2	0.95	0.91	<0.001	<0.001
Being overly critical of the resident (n11)	23	23	0	0.87	0.77	<0.001	—
Adopting an intimidating demeanor, or treating the resident in an overly rude, condescending, or abrasive manner (n12)	35	32	3	0.88	0.78	<0.001	<0.001
Speaking ill of other residents who are not present (n13)	6	6	0	0.89	0.80	0.004	—
Negative, not otherwise specified	12	6	6	0.17	0.09	0.553	—

Negative themes, total number of evaluations containing each theme, distribution of themes among above- and below-average evaluations, interrater reliability using Cohen Kappa and ICC (average measures), distribution using two-sided Fisher exact test, and logistic regression significance (when applicable).
ICC = Intraclass Correlation Coefficient.

Council for Graduate Medical Education,[1] expert panels,[17] faculty,[18] and trainees,[8,18] alike. The importance of feedback has only recently been recognized, having been rarely mentioned in descriptions of the ideal teacher just a generation ago.[19] Paradoxically, by providing such feedback, teaching faculty may lower their teaching scores, as compared with merely providing praise.[20] At the same time, however, being merely personable, collegial, and respectful was not sufficient to ensure a high-scoring evaluation. The single most common theme found in above-average evaluations related to the faculty member spending an adequate time teaching.

Negative themes had a high degree of reliability and were significantly associated with below-average evaluations. This establishes these themes as possible causes for the low teaching scores. Many of the negative themes represent behaviors that are the negative corollary of the exemplary behaviors in the positive themes. Avoiding these behaviors may provide an avenue for faculty members to improve trainee satisfaction with their teaching and supervision, and subsequently improve their own teaching scores. Our residents frequently request more intraoperative teaching, and this was reflected in the comments we studied. Similarly, in a similar recent study of free-text comments, the most common "area for improvement" comment was a request for more teaching.[21] Several of our negative themes may be easy for clinical teachers to manage, such as avoiding derogatory comments about residents, or taking the time to explain clinical decision-making. In a recent contextual analysis of clinical teaching faculty evaluations, Myers[21] concluded that written comments "… seem unlikely to provide faculty with substantive feedback." In contrast, our study reveals pointed and constructive criticism in low-scoring evaluations, with a high degree of reliability and a significant and independent correlation with having a low teaching score. This difference in conclusion may stem from our focus on high- and low-scoring evaluations, which may contain more substantive content, as well as our use of more evaluations over a longer period of time.

Taken together, our results have implications for clinical teachers. The numeric teaching score has little value independent of the associated comments, at least for above- or below-average evaluations. Most faculty members included in this study received both above- and below-average evaluations (fig. 1), and most residents submitted both above- and below-average evaluations (fig. 2). Though periodic evaluations may have a positive effect on teaching scores, a recent review

Table 5. Summary: Key Recommendations Based on Above- and Below-average Evaluations

		Theme
Teaching	p1	Support teaching with primary literature
	p2, n1, n6	Explain your clinical decision-making
	p3, p4, p6, n2	Make an effort to spend additional time teaching, make teaching a priority
	p7, p8	Make clinically relevant teaching a priority
Supervision	p9, n4, n5, n6, n7	Give autonomy as appropriate; maintain appropriate supervision always
	p12	Challenge your residents to a higher level of performance
	p11	Be patient and supportive while teaching a new procedure
	p13	Encourage the use of new methods or procedures
	n8	Maintain your clinical practice (both skills and knowledge)
Feedback	p1, n9	Give clear, constructive, and developmental feedback
Interpersonal	p15, n12	Treat the resident collegially and respectfully
	n11, n13	Be gentle when providing criticism; never criticize a resident who isn't present
	p15, n10, n11, n12	Avoid displays or expressions of frustration, anger, or impatience; provide criticism in an appropriate manner, at the appropriate time

concluded that evaluations were insufficient as a means to improve teaching effectiveness.[22] For low-performing faculty, receiving low-scoring evaluations may in fact have a negative effect on teaching performance.[5] Moreover, as with any individual,[23] faculty members may have limited insight into their own teaching effectiveness.[24] Use of observation, mentorship, and other faculty development tools have been demonstrated to improve teaching performance, with some having positive effects lasting for years.[22]

A particular strength of this study is the combination of a contextual analysis (recurring themes found in free-text comments) with above- or below-average evaluations. Contextual analyses have been done previously establishing different domains of the ideal clinical teacher.[10,21] To date, however, none of these themes had been demonstrated to correlate with positive evaluations or high teaching scores. This study demonstrates correlations with key behaviors in each of the four domains, both positively and negatively, and thus, provides some validation for each domain. It is important to state that our results are correlations, and thus, we cannot state cause and effect. These results are best characterized as descriptive of our resident-clinical teacher interactions.

Findings from this study may be limited by selection bias because we excluded evaluations that were not associated with above- or below-average teaching scores. This allowed for efficient detection of positive and negative teaching themes; however, further study of the themes found in average-scoring evaluations might help elucidate more nuanced themes. Additionally, the potential for reporting bias may have resulted from our subjective designation of comments into different categories, our use of a single individual (Dr. Baker) to create the initial set of themes, and the lack of specific definitions for several of our themes. Creating the theme list using the Delphi method may have been preferable; however, the results of independent logistic regression models for each rater were comparable, supporting the overall theme construction. They also had extremely strong correlations with the appropriate category of above- or below-average teaching scores. The designated themes captured the majority of the comments in these evaluations because less than half included "positive NOS" or "negative NOS." Many of these NOS comments had no developmental or meaningful content; many simply stated "great teacher." Some of the residents submitting evaluations were excluded from our study database due to the tendency to give the same teaching score to all faculty members, precluding the ability to separate above- or below-average scores. Our analysis did not take into account the hierarchical nature of the data structure, and we treated each evaluation as an independent measure. When the same resident evaluates the same faculty member on more than one occasion, the evaluations are unlikely to be independent. This may have inflated the statistical significance of some themes. However, this would not have any effect on the interrater reliability. Further study is warranted using independent samples from multiple institutions. Finally, we reiterate the correlative nature of our results, which precludes our ability to state cause and effect.

This study is predicated on the idea that residents submitting evaluations accurately identify and report valid reasons for the high or low teaching score that they assign. Using learner evaluations to assess the quality of teaching has long been a subject of debate.[16] Given the lack of a "definitive standard" for assessing clinical instruction, trainee feedback remains the main avenue for assessment. Recent advances include validated instruments,[25] specific benchmarks set by expert panels,[17] and focus groups to help define "better teaching."

This study found specific positive and negative themes in the comments section of resident evaluations of clinical teachers and related these themes to above- and below-average teaching scores. This study provides an association between above-average teaching scores and the behaviors associated with excellent teaching. Conversely, this study provides an association between below-average teaching scores and the behaviors associated with below-average teaching. These themes are recast as recommendations in table 5.

Acknowledgments

The authors thank Eleanor Cotter, A.S. (Department of Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts), for entering the evaluations and comments, and each of the residents, who took the time to evaluate and provide feedback to their clinical teachers, and Kevin Tremper, M.D., Ph.D. (Department of Anesthesiology, University of Michigan, Ann Arbor, Michigan).

Support was provided solely from institutional and/or departmental sources.

Competing Interests

The authors declare no competing interests.

Correspondence

Address correspondence to Dr. Baker: Department of Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, Gray-Bigelow 444, 55 Fruit Street, Boston, Massachusetts. khbaker@partners.org. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. Anesthesiology's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

References

1. The Accreditation Council for Graduate Medical Education. "Common Program Requirements", "Effective: July 1, 2011" Edition. The Accreditation Council for Graduate Medical Education, 2007
2. Atasoylu AA, Wright SM, Beasley BW, Cofrancesco J Jr, Macpherson DS, Partridge T, Thomas PA, Bass EB: Promotion criteria for clinician-educators. J Gen Intern Med 2003; 18:711–6
3. Beasley BW, Wright SM, Cofrancesco J Jr, Babbott SF, Thomas PA, Bass EB: Promotion criteria for clinician-educators in the United States and Canada. A survey of promotion committee chairpersons. JAMA 1997; 278:723–8
4. Dent MM, Boltri J, Okosun IS: Do volunteer community-based preceptors value students' feedback? Acad Med 2004; 79:1103–7
5. Litzelman DK, Stratos GA, Marriott DJ, Lazaridis EN, Skeff KM: Beneficial and harmful effects of augmented feedback on physicians' clinical-teaching performances. Acad Med 1998; 73:324–32
6. Baker K: Clinical teaching improves with resident evaluation and feedback. Anesthesiology 2010; 113:693–3
7. Maker VK, Curtis KD, Donnelly MB: Faculty evaluations: Diagnostic and therapeutic. Curr Surg 2004; 61:597–1
8. Boor K, Teunissen PW, Scherpbier AJ, van der Vleuten CP, van de Lande J, Scheele F: Residents' perceptions of the ideal clinical teacher—A qualitative study. Eur J Obstet Gynecol Reprod Biol 2008; 140:152–7
9. Kisiel JB, Bundrick JB, Beckman TJ: Resident physicians' perspectives on effective outpatient teaching: A qualitative study. Adv Health Sci Educ Theory Pract 2010; 15:357–68
10. Ullian JA, Bland CJ, Simpson DE: An alternative approach to defining the role of the clinical teacher. Acad Med 1994; 69:832–8
11. Masunaga H, Hitchcock MA: Residents' and faculty's beliefs about the ideal clinical teacher. Fam Med 2010; 42:116–20
12. Kornell N, Bjork RA: Learning concepts and categories: Is spacing the "enemy of induction"? Psychol Sci 2008; 19:585–92
13. Cicchetti DV: Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment 1994; 6:284–90
14. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49:1373–9
15. Steiner IP, Franc-Law J, Kelly KD, Rowe BH: Faculty evaluation by residents in an emergency medicine program: A new evaluation instrument. Acad Emerg Med 2000; 7:1015–21
16. McKeachie WJ: Student ratings: The validity of use. American Psychologist 1997; 52:1218–25
17. Yeates PJ, Stewart J, Barton JR: What can we expect of clinical teachers? Establishing consensus on applicable skills, attitudes and practices. Med Educ 2008; 42:134–42
18. Buchel TL, Edwards FD: Characteristics of effective clinical teachers. Fam Med 2005; 37:30–5
19. Sutkin G, Wagner E, Harris I, Schiffer R: What makes a good clinical teacher in medicine? A review of the literature. Acad Med 2008; 83:452–66
20. Boehler ML, Rogers DA, Schwind CJ, Mayforth R, Quin J, Williams RG, Dunnington G: An investigation of medical student reactions to feedback: A randomised controlled trial. Med Educ 2006; 40:746–9
21. Myers KA, Zibrowski EM, Lingard L: A mixed-methods analysis of residents' written comments regarding their clinical supervisors. Acad Med 2011; 86(10 suppl):S21–4
22. Steinert Y, Mann K, Centeno A, Dolmans D, Spencer J, Gelula M, Prideaux D: A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8. Med Teach 2006; 28:497–26
23. Kruger J, Dunning D: Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 1999; 77:1121–34
24. Claridge JA, Calland JF, Chandrasekhara V, Young JS, Sanfey H, Schirmer BD: Comparing resident measurements to attending surgeon self-perceptions of surgical educators. Am J Surg 2003; 185:323–7
25. Lombarts KM, Bucx MJ, Arah OA: Development of a system for the evaluation of the teaching qualities of anesthesiology faculty. Anesthesiology 2009; 111:709–16